Семантическая сегментация веб-гипертекста на основе дискретных математических моделей

В.С. Салин, С.В. Папшев
Саратовский государственный технический университет
им. Ю.А. Гагарина
salinvs@gmail.com, spapshev@list.ru

Аннотация

Основой различных современных подходов к эффективному поиску и извлечению информации в Интернете является построение семантического информационного слоя над гипертекстовым массивом. В данной статье рассматриваются подходы к семантическому насыщению гипертекста на основе дискретных математических моделей. Базой для решения задачи придания семантики гипертексту служит кластеризация веб-документов по различным признакам, включая их семантическую близость. В данной работе предложен метод построения семантических кластеров в гипертекстовой структуре веб-сайта на основе учета статистики переходов пользователей между узлами. Кластеризация документов по отслеживаемым маршрутам пользователей применяется к графовой модели гипертекстовой структуры веб-сайта. Предлагаемый метод семантической кластеризации основан на алгоритме кластеризации взвешенного графа BorderFlow. Для автоматизированного построения графовой модели гипертекста, а также применения к нему разработанных методов кластеризации, спроектирован и реализован программный комплекс. В области семантического веба, результаты могут использоваться для программного семантического анализа веб-документов. В веб-разработке и проектировании, результаты исследования помогают эффективно решать задачу построения адаптивной навигации, а также помогают при реинжиниринге веб-сайта и оптимизации его логической структуры для пользователей.

Ключевые слова: гипертекстовая структура; семантическая кластеризация; конечный автомат; граф; веб-аналитика

1. Иследования в области кластеризации гипертекста

При работе со значительными объемами информации Интернета одной из актуальных проблем является поиск релевантной запросу размещенной на веб-страницах информации. При этом возникают задачи анализа семантики веб-

страниц и структуризации информации. Важным этапом при их решении является кластеризация обрабатываемых документов, которая позволяет выявлять группы семантически похожих документов.

Кластеризация веб-страниц по тематике актуальна в различных приложениях поиска и интеллектуального анализа данных, таких как распознавание шаблонов, извлечение ключевых слов. Группировка семантически связанных страниц лежит в основе рекомендательных алгоритмов и адаптивных интерфейсов, которые могут «подсказывать» пользователю наиболее релевантные для него страницы. Определение кластеров близких по теме страниц используется также в веб-аналитике и поисковой оптимизации, разработке и реинжиниринге веб-сайтов.

Для решения задач кластеризации наиболее просто применить методы непосредственного определения ключевых слов и понятий текста. Традиционно, при данном подходе объекты текста описываются характеристическим вектором. Векторы могут быть интерпретированы как точки в многомерном пространстве. В классических методах кластеризации используется некоторая метрика расстояния, например, косинус угла между такими векторами. Реализация таких методов имеет определенные сложности: необходимость предварительной индексации, высокое время выполнения при больших объемах текстов, с другой стороны – необходимость в наличии минимального объема текста с релевантными ключевыми словами.

Альтернативой текстовой кластеризации, в контексте задачи кластеризации гипертекстовых документов, является кластеризация, построенная на основе информации, содержащейся в гипертекстовой структуре. Данная информация отличает гипертекстовые документы от простых, текстовых документов и может включать в себя: информацию о внешних гиперссылках, информацию о внутренних связях в гипертекстовой структуре, данные о статистике переходов пользователя по элементам гипертекста и т.д.

Впервые задача кластеризации была озвучена в 30-х годах XX века, в работах Р. Триона [1] и Р. Б. Каттела [2] и имела приложение в областях антропологии и психологии. Широкое распространение кластеризация данных получила с развитием подходов к интеллектуальному анализу данных ближе к концу XX века, сформировав отдельное направление кластерного анализа. Теоретические и методологические основы на данном направлении заложены в результатах исследований многих отечественных и зарубежных авторов, включая Б. Дюрана, П. Оделла [3], И. Д. Манделя [4], С. А. Айвазяна [5], Д. С. Хайдукова [6] и других.

В области информационных технологий актуальна задача кластеризации документов, или семантической кластеризации. Применительно к кластеризации веб-документов, стоит отметить обзор К. Карпинето и др. [7], в котором авторы рассматривают 14 алгоритмов текстовой кластеризации для решения задачи группировки результатов поиска. Однако рассмотренные ими подходы хоть и отличаются от классических центроидных алгоритмов кластеризации документов, однако ориентированы на рассмотрение веб-документа как текстового объекта и не учитывают его внутренних гипертекстовых свойств.

Применение математических моделей при анализе гипертекстовой структуры для различных задач позволяет использовать соответствующий выбранной модели математический аппарат с учетом внутренних свойств гипертекста. В 90-х годах существенный вклад в разработку математических моделей гипертекста внесли Д. Скоттс и Р. Фурута [8–11], последовательно предложив целый ряд моделей, включая графовую модель, модель сети Петри, автоматную модель, решетчатую модель. Д. Б. Ланж предложил объектно-ориентированный подход к моделированию [12], Ф. Дж. Халасц и М. Д. Шварц разработали модель Декстера для гипертекста [13].

Вместе с появлением концепции Web 2.0 и развитием технологий разработки веб-сайтов, усложнением их структуры, стремительным увеличением объемов информации исследователи вернулись к вопросу анализа поведения пользователя в гипертексте, проблемам построения удобной, визуальной, адаптивной навигации и пользовательского интерфейса. Подробный обзор работ, посвященных анализу навигационного поведения пользователей для решения задач построения пользовательских интерфейсов и навигации в гипертексте, провели авторы В. Холинк, М. ван Сомерен, Б. Дж. Вилинга [14]. В частности, исследования и разработка семантической и адаптивной моделей навигации отражены в работах авторов Б. Смита и П. Коттера [15], М. Дж. Паззани и Д. Биллсуса [16], М. Перковиц и О. Этциони [17], В.С. Салина, С.В. Папшева и А.А. Сытника [18]. Отдельно стоит отметить исследования Г. Бейдун и Р. Кульчитски [19], в котором авторы предлагают применять при построении навигационной модели в гипертексте статистику переходов пользователей, в то время как традиционно статистика переходов и посещений веб-страниц использовалась для решения задач веб-аналитики (А. Кошик [20]). В смежной области реинжиниринга веб-сайтов поведение пользователей предлагалось взять за основу в работе [21].

Дальнейшее развитие концепции Web 2.0 и ее трансформация в Web 3.0 сместили акцент с пользовательских веб-интерфейсов на программные интерфейсы доступа к данным в Вебе и автоматизированного анализа их семантики (Н. Шедболт, В. Холл, Т. Бернерс-Ли [22]). Как следствие, еще больше внимания стало уделяться методам обработки текстов естественного языка и интеллектуального анализа данных для извлечения семантики веб-страниц. При этом чаще всего веб-документы анализируются как простые текстовые документы, а свойства самого гипертекста не используются, равно как и не используется статистика переходов в нем.

Исходя из результатов проведенного анализа исследований в областях кластерного анализа, интеллектуального анализа данных, моделирования гипертекста, веб-аналитики и семантического веба, в данной статье предлагается использование статистики переходов в гипертекстовой структуре и разработка на данной основе моделей и методов семантической кластеризации данной структуры. В работе предлагается метод семантической кластеризации с использованием статистики переходов пользователей по гиперссылкам, в частности, на базе существующего метода графовой кластеризации разработан метод семантической кластеризации по статистике переходов. На основе данных моделей и методов реализован программный комплекс для решения задачи кластеризации, а также представлены практические рекомендации по

применению результатов исследования в решении прикладных задач в области поиска и анализа данных, веб-разработки, веб-аналитики и семантического веба.

2. Семантическая сегментация веб-гипертекста

2.1. Математические модели для решения задач кластеризации

Основой для представления гипертекстовых данных для последующей кластеризации может служить одна из математических моделей, все они из класса так называемых дискретных моделей. В зависимости от целей и применяемого метода используют модель в виде автомата, сети Петри или графа.

Кластеризация графов отличается от кластеризации объектов в многомерном пространстве, так как использует не метрику расстояния между объектами, а степень связности узлов графа, на основе которой определяется сходство между объектами (узлами). Достаточно удаленные друг от друга объекты могут не сравниваться, что делает использование данной модели более эффективным по сравнению с обычной кластеризацией, основанной на метрике расстояния.

Для решения задачи семантической кластеризации, гипертекстовую структуру можно рассматривать совместно с информацией о переходах, совершенных пользователями данной структуры от одного узла к другому в течение определенного времени. В соответствие с этим гипертекстовой структуре и статистике по совершенным в ней переходам можно поставить в соответствие математическую модель в виде взвешенного графа.

В общем случае, гипертекст может быть представлен как $H = \{P, L\}$, где $P = \{p_1, \ldots, p_n\}$ – множество гипертекстовых документов, $L = \{l \mid \exists p_1, p_2 \in P : p_1, p_2 \in l(p_1, p_2)\}$ – множество гиперссылок между ними.

Тогда формальная постановка задачи семантической кластеризации гипертекстовых документов будет выглядеть следующим образом. На множестве объектов – гипертекстовых документов $P = \{p_1, \ldots, p_n\}$ задана функция расстояния между объектами $\rho_{sem}(p, p')$. Требуется разбить исходное множество объектов на подмножества – кластеры, чтобы каждый кластер содержал объекты, близкие по метрике ρ_{sem}, а объекты разных кластеров максимально отличались по той же метрике. В результате каждому $p_i \in P$ приписывается номер $y_i \in Y$. Функция $f_{sem}: P \to Y$ называется функцией семантической кластеризации, а ее конкретная реализация зависит от выбранного метода кластеризации.

Будем учитывать поведение пользователей данной гипертекстовой структуры и рассматривать её совместно с маршрутами пользователей в ней. Предположим, что для страниц $P' \subset P$ за временной интервал ΔT известно множество маршрутов переходов $R = \{r_i\}$:

$$r_i = (p'_m, \ldots, p'_n),\ p'_j \in P',\ j \in \mathbb{N}$$

В таком случае функция семантической кластеризации должна учитывать не только данные о самой гипертекстовой структуре, но и о маршрутах в ней: $f_{sem}(H,R)$.

Очевидным образом гипертекстовой структуре $H = \{P, L\}$ может быть однозначно поставлена в соответствие математическая модель в виде ориентированного невзвешенного графа $G = \{V, E\}$, в котором $V = P$ – вершины графа, соответствующие гипертекстовым документам, $E = L$ – ребра графа, соответствующие гиперссылкам между документами.

Аналогично, маршруты пользователей r_i могут быть представлены графами $G_i' = \{V_i', E_i', W_i'\}$, где $V_i' \subset V$ – посещенные вершины, $E_i' \subset E$ – совершенные переходы между ними, W_i' – веса в графе маршрутов, где вес каждого ребра соответствует количеству переходов по гиперссылке. Также для каждого графа отмечена корневая вершина $v_{i\,0}'$. Все множество маршрутов за промежуток ΔT выглядит как $R = \{G_i'\}$.

Таким образом, задача семантической кластеризации гипертекстовой структуры сводится к задаче кластеризации соответствующего графа, а функция семантической кластеризации примет вид:

$$f_{sem}(G,R): V \to Y, R = \{G_i'\}$$

где G_i' – маршруты пользователей гипертекстовой структуры.

Обратным преобразованием кластеров – подмножеств вершин графа – к множеству гипертекстовых документов будет получено разбиение гипертекста на кластеры.

2.2. Метод семантической кластеризации гипертекстовой структуры на основе статистики переходов

Разработанный метод семантической кластеризации гипертекстовой структуры основывается на графах маршрутов пользователей в данной структуре. Графы маршрутов соответствуют последовательностям переходов пользователей между узлами гипертекста.

Пусть маршруты пользователей r_i представлены орграфами $G_i' = \{V_i', E_i', W_i'\}$. Для произвольных двух графов вводится понятие общей начальной подпоследовательности переходов $\varepsilon(G_j', G_k'), |\varepsilon| \in [0 \dots n], n \in \mathbb{N}$. Она содержит общую последовательность ребер, начинающихся с корня данных графов, а ее длина $|\varepsilon|$ показывает, сколько общих ребер от корня есть в данном графе. В случае $v_{i\,0}' \neq v_{j\,0}', |\varepsilon| = 0$.

Из всего множества графов маршрутов формируются группы графов, для которых $|\varepsilon| \geq \varepsilon_0$. В самом простом случае $\varepsilon_0 = 1$.

В каждой полученной группе l производится слияние матриц смежности графов, входящих в данную группу. Для чего сперва объединяются множества вершин графов в группе: $V_l = \cup\{V_1, \dots, V_m\}$, после чего строится матрица $M_l = N \times N$, где $N = |V_l|$. Общий вид матрицы будет выглядеть следующим образом:

$$M_l = \begin{bmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & \ddots & \vdots \\ c_{n1} & \cdots & c_{nn} \end{bmatrix}, \; c_{jk} \in \mathbb{N}$$

Каждый элемент матрицы вычисляется по формуле:

$$c_{jk} = \sum_{i=1}^{m} w\left(e(v_j, v_k)\right), w \in W_i$$

Полученной матрице будет соответствовать взвешенный ориентированный граф $G_l = \{V_l, E_l, W_l\}$. Чтобы исключить случайные переходы, часть вершин исходного множества V_l отсеивается по весам, большим порогового значения w_0:

$$X_l = \{v \in V_l | \forall u \in V_l : w(e(v,u)) > w_0,\ w \in W_l, e \in E_l\}$$

Повторив действия для каждой полученной группы l, получается в результате множество вершин $\{X_l\}$:

$$X_{st} = \{X_l\},\ X_l \subset V,\ l \in [1..t]$$

Данное множество будет являться результатом семантической кластеризации исходного множества вершин графа.

2.3. Метод кластеризации гипертекстовой структуры с использованием ее внутренних связей и статистики переходов по ним

Статистика переходов в гипертексте может также применяться для семантической кластеризации совместно с информацией о внутренних связях в гипертекстовой структуре. В рамках данного исследования, разработан метод семантической кластеризации основанный на существующем алгоритме кластеризации графа BorderFlow, веса для которого формируются исходя из статистики переходов.

Метод кластеризации BorderFlow определен для взвешенного орграфа $G = \{V,\ E,\ W\}$. В его основе лежит концепция объединения в кластеры узлов, имеющих веса ребер внутри кластера большие, чем снаружи. Чтобы применить BorderFlow к графовой модели гипертекста, необходимо определить множество весов W для него. В данном случае, статистика переходов в гипертексте является источником данных для составления весов в графе.

Основываясь на всех графах маршрутов пользователей, полученных за выбранный интервал времени, составляется общая матрица $M = N \times N$, учитывающая количество переходов во всей гипертекстовой структуре. Размерность матрицы $N = |\cup\{V_1, \dots,\ V_m\}|$, где V_i – вершины, посещенные в рамках *i*-го маршрута.

$$M = \begin{bmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & \ddots & \vdots \\ c_{n1} & \cdots & c_{nn} \end{bmatrix},\ c_{jk} \in \mathbb{N}$$

Каждый элемент матрицы вычисляется по формуле:

$$c_{jk} = \sum_{i=1}^{m} w\left(e(v_j, v_k)\right), w \in W_i$$

Полученной матрице будет соответствовать взвешенный ориентированный граф $G' = \{V', E', W'\}$. Сопоставляя данный граф с исходным графом гипертекста $G = \{V,\ E\}$, важно отметить, что $V' \subseteq V, E' \subseteq E$.

К полученному графу $G' = \{V', E', W'\}$ применяется алгоритм кластеризации BorderFlow, результаты его работы трактуются как результаты семантической кластеризации гипертекста.

2.4. Программный комплекс для автоматизированного построения модели гипертекста и применения методов кластеризации с учетом статистики переходов

Для автоматизированного построения графовой модели гипертекста, а также применения к нему разработанных методов кластеризации, спроектирован и реализован программный комплекс. Программная система семантической кластеризации включает в себя четыре основных компонента, показанные на диаграмме (Рисунок 1). Первый программный компонент (на диаграмме обозначен буквой A) ответственен за построение модели веб-сайта. В его работу входит сканирование веб-сайта с целью восстановления гипертекстовой структуры, которая далее представляется в виде графовой модели.

Второй компонент (B) занимается хранением модели просканированного гипертекста. В нем сохраняется вся необходимая информация об узлах (URI и заголовок веб-страницы) и связях между ними. При сохранении, графовая модель гипертекста соответствующим образом отображается на ER-модель хранилища.

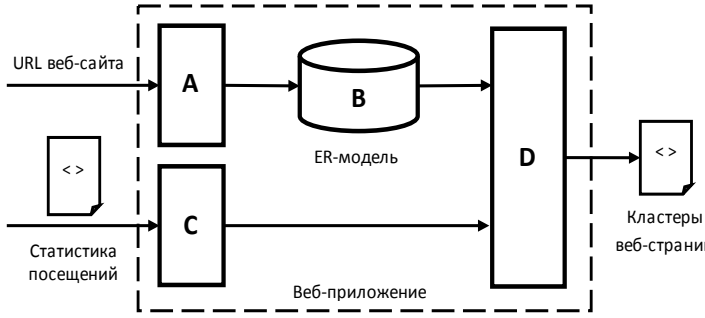

Рис. 1. Диаграмма компонентов программной системы для семантической сегментации

Компонент (C) предназначен для получения информации о посещении веб-сайта. Являясь адаптером к системам сбора статистики, таким как Google Analytics или Яндекс.Метрика , компонент запрашивает собранную статистику посещений узлов сайта и переходов между ними. Данная информация используется в дальнейшем для семантической сегментации веб-сайта.

Четвертый компонент (D) выполняет основную функцию семантической сегментации на графовой модели веб-сайта с использованием статистики переходов. Графовая модель получается обратным преобразованием из ER-модели хранилища, информация о статистике посещений поступает из компонента (C). Результатом работы компонента (D), как и результатом работы

всей системы, являются кластеры веб-страниц сайта, представленные в файле легковесного CSV-формата.

Согласно рассмотренной архитектуре, программная система реализована как веб-приложение на платформе Java (компоненты A, C, D) с использованием СУБД MySQL в качестве хранилища (B). При реализации, помимо стандартных средств платформы Java SE использовались сторонние библиотеки, в частности, для сканирования веб-сайтов и построения их гипертекстовой структуры, формирования файла полученных кластеров, подключения и работы с системой сбора статистики, подключения к БД MySQL.

2.7. Области приложения методов семантической кластеризации

Часть разработанного программного комплекса, ответственная за построение математической модели и применения к ней методов семантической кластеризации, оформлена в виде программной библиотеки и может быть подключена для использования через программный интерфейс в решении прикладных задач в смежных областях. Для соответствующих областей, подготовлены практические рекомендации к использованию разработанных и реализованных методов.

В области семантического веба, разработанные методы могут применяться в задачах программного анализа данных и выявления семантики веб-документов. В частности, семантическая кластеризация веб-документов позволяет определять ресурсы, схожие по смыслу с заданным, не прибегая к текстовому анализу содержимого. Кроме того, применение методов статистического анализа текста к кластерам семантически близких документов помогает в задаче выявления фактов и терминов. Отсутствие необходимости в непосредственном анализе текстового содержимого веб-документов позволяет применять разработанные методы в тех случаях, когда традиционные методы текстовой кластеризации не применимы или имеют существенные сложности.

В разработке и проектировании веб-сайтов, разработанные модели и методы могут применяться для решения задач реинжиниринга веб-сайта, реализации рекомендательных механизмов, оптимизации логической структуры веб-сайта с учетом поведения пользователей. Задача построения адаптивной навигации на основе рекомендаций пользователю релевантных страниц является особенно актуальной для большинства веб-сайтов концепции Web 2.0. Текущие варианты ее решения основываются либо на методах текстовой кластеризации, либо на ручном экспертном присвоении меток (тегов) кластеров конкретным веб-документам и часто сталкиваются с ограничениями обоих подходов. Применение семантической кластеризации на основе статистики переходов пользователей не требует ни текстового анализа содержимого, ни ручной разметки документов тегами.

Разработанные методы могут применяться также при решении задачи семантической кластеризации при поиске веб-документов, встроенном в функционал веб-сайта. В этом случае семантическая кластеризация позволит сформировать кластеры близких по смыслу ресурсов, релевантных поисковому запросу. При этом для самой кластеризации не потребуется текстовый анализ всех веб-страниц, так как формирование будет производиться на основе статистики посещений.

3. Основные результаты исследования

В результате проведенного исследования разработана математическая модель семантической кластеризации гипертекстовой структуры, а также методы кластеризации данной гипертекстовой структуры, использующие статистику переходов и внутреннюю структуру гиперссылок.

Предложенная математическая модель гипертекстовой структуры имеет расширенный набор свойств по сравнению с существующими моделями (графовыми, автоматными). Данные свойства позволяют учитывать статистику переходов между гипертекстовыми документами, а, следовательно, появляется возможность применять новые методы семантической кластеризации, основанные на статистике переходов. В то же время, модель не использует текстовое содержание документов и может быть использована в тех случаях, когда анализ текста затруднен или невозможен.

Разработанный метод кластеризации гипертекстовых документов основан на использовании статистики переходов между документами внутри гипертекстовой структуры и не требует полнотекстовой индексации и последующего поиска по индексу, что отличает от традиционных методов текстовой кластеризации и повышает его эффективность. В отличие от методов, построенных на анализе внешних гиперссылок, метод не требует наличия существенной базы подобных гиперссылок и определения их тематики.

Предложен метод семантической кластеризации, в котором статистика переходов используется для определения весов в алгоритмах графовой кластеризации, что позволяет эффективно применять данные методы для семантической кластеризации гипертекста.

На основе разработанных методов разработана программная система для автоматизированного построения модели гипертекстовой структуры и применения к ней разработанных методов кластеризации на основе статистики переходов. Данная система может применяться как инструмент для решения прикладных задач в областях проектирования, разработки и реинжиниринга веб-сайтов, веб-аналитике, поиске и интеллектуальном анализе данных, в задачах семантического веба.

В области семантического веба, результаты могут использоваться для программного семантического анализа веб-документов, для выявления фактов в кластере веб-страниц, установки семантических связей между ними. В веб-разработке и проектировании, результаты исследования помогают эффективно решать задачу построения адаптивной навигации, а также помогают при реинжиниринге веб-сайта и оптимизации его логической структуры для пользователей.

Литература

[1] Tryon R.C. Cluster analysis // London: Ann Arbor Edwards Bros, 1939. — 139 p.

[2] Cattell R. B. The description of personality: Basic traits resolved into clusters // Journal of Abnormal and Social Psychology. 1943. Vol. 38. P. 476–506

[3] Дюран Б., Оделл П. Кластерный анализ. — М.: Статистика, 1977. — 128 с.

[4] Мандель И. Д. Кластерный анализ. — М.: Финансы и статистика, 1988. — 176 с
[5] Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: Классификация и снижение размерности. — М.: Финансы и статистика, 1989. — 607 с.
[6] Хайдуков Д. С. Применение кластерного анализа в государственном управлении// Философия математики: актуальные проблемы. — М.: МАКС Пресс, 2009. — 287 с.
[7] Carpineto C., Osiński S., Romano G., Weiss D. A survey of Web clustering engines // ACM Computing Surveys (CSUR). 2009. Vol. 41, №3.
[8] Stotts D., Furuta R. Adding browsing semantics to the hypertext model // DOCPROCS '88 Proceedings of the ACM conference on Document processing systems. 1988. P. 43-50
[9] Stotts D., Furuta R. Petri-net-based hypertext: document structure with browsing semantics // ACM Transactions on Information Systems (TOIS). 1989. Vol. 7, №1, P. 3-29
[10] Stotts D., Furuta R. The trellis hypertext reference model // Computer science technical report series. University of Maryland. 1990. 11 p.
[11] Stotts D., Furuta R. Hyperdocuments as Automata: Verification of Trace-based Browsing Properties by Model Checking // ACM Transactions on Information Systems (TOIS). 1998. Vol.16, №1. P. 1-30
[12] Lange D.B., A Formal Approach to Hypertext using Post-Prototype Formal Specification // VDM '90 Proceedings of the Third International Symposium of VDM Europe on VDM and Z - Formal Methods in Software Development. 1990. P. 99-121
[13] Frank G. Halasz, Mayer D. Schwartz, The Dexter hypertext reference model // Communications of the ACM. 1994. Vol. 37, №2. P. 30-39
[14] Hollink V., Someren M., Wielinga B.J. Navigation behavior models for link structure optimization // User Modeling and User-Adapted Interaction. 2007. Vol. 17, №4, P. 339-377
[15] Smyth, B., Cotter, P.: Intelligent Navigation for Mobile Internet Portals Proceedings of the 18th International Joint Conference on Artificial Intelligence, AI Moves to IA: Workshop on Artificial Intelligence, Information Access, and Mobile Computing, Acapulco, Mexico. 2003
[16] Pazzani M.J., Billsus D. Adaptive web site agents // Journal of Agents and Multiagent systems. 2002. Vol. 5 №2 P. 205-218
[17] Perkowitz M., Etzioni O. Adaptive Web Sites: Automatically Synthesizing Web Pages. // Proceedings of the Fifteenth National Conference on Artificial Intelligence. 1998
[18] Салин В. С., Папшев С. В. и Сытник А. А. Об одном методе синтеза семантической структуры веб-сайта // Вестник Саратовского государственного технического университета. Саратов, 2011 г. - 60. - С. 199-202.
[19] Beydoun G. Formal concept analysis for an e-learning semantic web // Expert Systems with Applications. Elsevier, 2009. Vol. 36. P. 10952-10961.
[20] Кошик А. Веб-аналитика 2.0 на практике. Диалектика, 2011.

[21] Салин В. С., Папшев С. В. и Сытник А. А. About a Method of Educational Web Resources Optimization // International Conference on Engineering Education and Research (iCEER 2013). Маракеш, 2013. С. 829-835.
[22] Shadbolt N., Hall W., Berners-Lee T. The Semantic Web Revisited // IEEE Intelligent Systems. 2006.

Semantic Segmentation of Hypertext Based on Discrete Mathematical Models

V. Salin, S. Papashev
The Yuri Gagarin State Technical University of Saratov

The issue of different modern methods for effective information search and data extraction is semantic layer construction up to hypertext information massive. We consider some approaches to semantic enhancing for hypertext based on different mathematical models. We see the basis for adding semantics to hypertext in clustering of web documents on various attributes, including its semantic similarity. In this paper we present the method for constructing semantic clusters of hypertext website structure based on web-analytic results of user visits and transitions across hyperlinks. We apply documents clustering by user's routes to graph model of the hypertext structure. The method of semantic clusterization, proposed in the paper, is based on the algorithm BorderFlow for weighted graphs. To automate graph model construction for hypertext, and to apply to it developed methods the special software was developed. The research results may be used for semantic analysis of web-documents. For web-design and developing purpose it may be applied to decide effectively the adaptive navigating task, and also help during web site reengineering and logic structure optimization process.

Keywords: hypertext structure, semantic clustering, finite automata, graph, web-analytics